Can We Fix the Security Economics of Federated Authentication?

Ross Anderson

There has been much academic discussion of federated authentication, and quite some political manoeuvring about 'e-ID'. The grand vision, which has been around for years in various forms but was recently articulated in the US National Strategy for Trustworthy Identities in Cyberspace (NSTIC), is that a single logon should work everywhere [1]. You should be able to use your identity provider of choice to log on anywhere; so you might use your driver's license to log on to Gmail, or use your Facebook logon to file your tax return. More restricted versions include the vision of governments of places like Estonia and Germany (and until May 2010 the UK) that a government-issued identity card should serve as a universal logon. Yet few systems have been fielded at any scale.

In this paper I will briefly discuss the four existing examples we have of federated authentication, and then go on to discuss a much larger, looming problem. If the world embraces the Apple vision of your mobile phone becoming your universal authentication device – so that your phone contains half-a dozen credit cards, a couple of gift cards, a dozen coupons and vouchers, your AA card, your student card and your driving license, how will we manage all this? A useful topic for initial discussion, I argue, is revocation. Such a phone will become a target for bad guys, both old and new. What happens when someone takes your phone off you at knifepoint, or when it gets infested with malware? Who do you call, and what will they do to make the world right once more?

Case 1 – SSO

Perhaps the oldest fielded example of federated authentication is corporate single sign-on. The goal of such systems is to enable a company's employees to log on to a diversity of internal applications with one password or token. I have been wrestling with such systems at various employers and consultancy clients since the 1980s. Even before the minicomputer ended the dominance of the corporate mainframe, employees faced multiple logons; a bank might have its branch accounting system running on top of MVS while its treasury systems ran on DB2 and its internal HR on top of something else again. The proliferation of Unix, Windows and Cloud systems has made life ever harder.

The main lesson is that even where all the users are the staff of a single company, which has unity of purpose, and the systems are purchased and maintained by a single IT organisation that tries hard to manage complexity in order to control costs, the battle for single sign-on is never won. There are always systems that just don't fit. Even in young high-tech firms with everyone trying to pull in the same direction – in short, where there are no security-economics issues of strategic or adversarial behaviour between firms – there are always new apps for which the business case is so strong that exceptions are made to the rules. This should warn us of the inherent limits of any vision of a universal logon working for all people across all systems everywhere.

Case 2 – SSL

The second cautionary tale comes from the proliferation of certification authorities (CAs). After SSL was adopted as the default mechanism for web authentication in the mid-1990s, and a handful of firms like Verisign and Baltimore cornered the market, people rightly objected. Things have since gone to the other extreme with hundreds of CAs having their certificates embedded in common browsers; and as Chris Soghoian has documented, many of these appear to have no function beyond enabling various police and intelligence services to perform silent meet-in-the-middle attacks on SSL sessions, and to covertly install surveillance software on people's machines [2].

Might this be fixed? At the authentication workshop following FC2011 I asked a panelist from the Mozilla Foundation why, when I updated Firefox the previous day, it had put back a certificate I'd previously deleted, from an organisation associated with the Turkish military and intelligence services. The Firefox spokesman said that I couldn't remove certificates – I had to leave them in but edit them to remove their capabilities – while an outraged Turkish delegate claimed that the body in question was merely a 'research organisation'. The Firefox guy then asked what sort of protocol might be appropriate for denying access to the certificate store in an open product; surely every organisation that meets publicly-stated norms should get in. They'd have to observe something bad happening before they yank a cert (though observing bad things is hard). This exchange shows how intractable the problem of global 'identity' provision has become. Perhaps it nudges us towards the relative naming in SPKI/SDSI [3], where there is no confusion between 'What Verisign calls Gmail' and 'What Tubitak calls Gmail'.

At least until then, global single sign-on will be made hazardous by government coercion. I may trust Google to be robust in resisting government attempts to read my Gmail, in that they will contest warrants presented by the police forces of countries in which I am neither resident nor located; but if Facebook and the people who issue driving licenses can also log me on to Gmail (and might thus be coerced to log on to Gmail as me), I have to think about their policies too.

Case 3 – 3DS

The third case study is 3D Secure, branded as MasterCard SecureCode and Verified by VISA, which enables you to use a password with your credit card to authenticate a payment at merchant websites. This was the banks' answer to the surge in cardholder-not-present fraud that followed the introduction of EMV payment cards in Europe.

We documented in [4] how 3DS has become the most widely-deployed single sign-on protocol in the world; despite having poor technical security, poor security usability and poor privacy, it has strong incentives for adoption – merchants who use it get their transactions treated as if the cardholder were present. This means lower interchange fees and less liability for chargebacks in the event of disputes; in effect, liability is passed to the cardholder. This is a good example of how bad engineering with strong adoption incentives can trump good engineering without them. (It may also be worth noting that

the hard issues such as enrolment and lost passwords are handled in 3DS by the card-issuing banks and their contractors.)

Case 4 – OpenID

The fourth case to consider is OpenID. After earlier attempts to set up proprietary global schemes (such as Microsoft Passport), a number of firms got together in 2007 to create a scheme under which each user chooses an 'identity service provider', and relying parties will redirect them to their chosen provider in order to authenticate them. Yet the uptake of this has been disappointingly slow. Kosta Beznosov and colleagues identify the problem as a lack of incentives for hosting and service companies, and a lack of demand from users [5]. Thus although there are over a billion OpenID-enabled accounts at large providers, there are few relying parties. His studies indicate that users are concerned about phishing; a quarter of them about single points of failure; that 40% are hesitant to consent to release of personal info when signing up with a relying party (which is actually the work of a separate protocol, OAuth, that's often bundled with OpenID to help websites collect personal information); and 36% said they wouldn't use single sign-on for critical sites like banking, or for valuable personal information, or on sites they did not believe to be trustworthy. In view of the above discussion, users are being rational in avoiding Open ID (and OAuth), as are many websites.

Mobile wallets

The application that motivates this paper is the mobile wallet. In Apple's vision of the future, your iPhone is not just a phone, address book, calendar, mail client and mobile browser, but also a wallet containing credit cards, store loyalty cards, gift cards, vouchers, your AA card and maybe even your driving license. The other smartphone technology companies – Microsoft, Google, and Blackberry – will no doubt follow suit, and we'll see wallets from other suppliers too.

The mobile wallet may become the new battlefield of information security. Until now, most malware has been written for Windows, and the bad guys monetise infected machines by selling them to botmasters. But price competition has been fierce; machines in the USA and Europe now sell for 13c each, while machines in Asia can be had for 3c. Mobile phones already offer a monetisation path through apps that silently call premium-rate lines, and we're beginning to see the malware industry producing a lot of these. Once phones contain wallets and malware can steal real money, the incentives will become stronger still. We may predict that whichever platform wins the current smartphone market race will become a major malware target – and perhaps even overtake Windows as *the* major malware target.

Also, once the average person's phone contains their money and all the other keys to their life, thefts of phones will become more common, and lost phones will become more serious. At present, about 2% of phones are lost or stolen each year; if at equilibrium we have a billion smartphone users (plus another two billion using less sophisticated phones, many in less developed countries) then every year twenty million people will suffer the

inconvenience and even anguish of having their digital lives lost or stolen. (We may also have to cope with a similar number of phones infected with malware, but in what follows I'll mostly discuss the lost-or-stolen case, as it's simpler.)

The trust architecture of a typical mobile wallet will have about four layers.

1. A secure element (SE), which is a smartcard chip, packaged along with the near-field communications (NFC) chip and mounted either on the phone's motherboard or (for older phones) as a plug-in accessory. The SE is available from several vendors; it contains implementations of payment protocols such as PayPass and EMV as well as Java Card, and will have one or more control applets as well as an applet implementing each credit card that the user loads on it.
2. The mobile phone itself, whether running Android, iOS or Windows, will have a wallet application that talks to the SE and which may in turn be called by other apps on the phone. With luck the wallet will provide a trustworthy user interface.
3. There will be an online service with which the wallet communicates and which provides logging, backup and other facilities. The mobile phone also goes online for similar services. However, while the online service provider may synchronise address-book data with the phone in the clear, it will not store clear values of the cryptographic keys that credit cards use to authenticate purchases. These keys will be encrypted in the SE for online backup.
4. Keys will be managed by a third party called a trust services manager (TSM), the online equivalent of the personalisation houses that at present contract with banks to issue EMV smartcards. Indeed the TSMs are likely to be the existing personalisation firms: their hardware security modules (HSMs) contain the key material needed to initialise cards, verify PINs, set up keys so that they or their customer banks can check the message authentication codes with which transactions are authorised, and so on.

When the user wants to install a credit card on her phone, she will call her bank which will identify her by whatever protocol it is comfortable with (which might range from asking her mother's maiden name, through sending a magic number in the post, to asking her to attend at a branch with her passport) and then authorise the TSM to load the payment card into the SE of her phone. (The SE comes pre-loaded with ignition key material that is available to the TSM, and that bootstraps this protocol.)

When she wishes to make a purchase, the SE talks to the merchant terminal via the NFC interface and the merchant then talks via its acquiring bank to the issuing bank. The phone might not be online – reception in stores can be poor, and the merchant terminal may be online only intermittently. But the phone will go online eventually once reception is restored, and then it can synchronise its transaction record with the cloud.

The final ingredient in the problem is that businesses naturally try as hard as they can to externalise costs by 'leveraging' the services of others. Online businesses find the cost of call centres onerous; one online bank, for example, has 3000 call centre staff but only 400 'proper' staff (of whom 150 are IT people), while a phone company reckons that each

call to its call centre costs $20. In the UK, call centres employ 3.5% of the workforce – over a million people – and there are still more in India [6]. So online businesses design their systems to minimise call-centre use. Each customer must typically designate an email address to be used to recover a lost password; and the large email service providers such as Gmail, Hotmail and Yahoo ask for a mobile phone number. But this game of pass-the-parcel has to end somewhere, and the problem becomes acute with mobile wallets: a mugger who steals my wallet thereby gets my phone, my email account and my money. When that happens, who am I to call?

A security-economics proposal

While the cost of call centres can be reduced by usable system design, it is unlikely that we can ever reduce customer contact to zero. Passing the cost (and liability) on like a hot potato to whoever will catch it cannot be the answer. Instead we should design the ecosystem so that each customer contact can be handled by whichever firm has the most to gain, or the most to lose.

If a mobile wallet is stolen, the parties with the most to lose are not the phone companies (to whom the marginal cost of minutes is near zero) but the banks. It is therefore rational for the customer to contact one her banks to cancel the credit cards in her mobile wallet. At present, if my physical wallet is stolen, I have to call all my banks one after the other, unless I am pre-registered with a card protection service. For example, one of my banks (Lloyds TSB) offers premium customers a free plan to notify all our card issuers following one phone call (less wealthy customers must pay £39.95 a year for the service). Mobile wallets allow such a service to be provided more cheaply: rather than notifying several other banks who then block payments through the interbank system, the contact bank can disable all the cards in the wallet at once by locking the SE. This is a better control point than doing something on the phone, in the cloud or in the bank transaction processing system. (It also helps in the case of phone malware – which is still rare, but is bound to increase, especially if careless phone vendors design wallets from which malware can easily steal money.)

Which bank should I call? The first part of my proposal is this. With a mobile wallet, unlike a leather wallet, the designer has to provide for a single credit or debit card to be the default. Wallets will no doubt have interfaces that allow the user to select the card of her choice for an in-store transaction, but for rapid transactions – tapping a subway turnstile or a parking meter – one card will be on top. And the privilege of being on top is immensely valuable. For example, there has been a long tussle between the US retail and banking industries over the approximately 1% higher charges levied for credit cards versus debit cards. Walmart would dearly love its debit card to be the default, while your bank would greatly prefer its credit card to be in that position. The natural way to align the incentives is for the default card issuer to be the firm you call to report a loss or theft.

Of course, a wallet provider cannot give a ‘revoke’ button to just anyone, because of the risk of abuse. The firm that wants its card on top will have to indemnify other card providers, and the phone company, against the consequences of improper revocation. But

the industry is used to this; my bank's scheme insures £1500 up till reporting and £75000 thereafter so long as the loss is reported within 24 hours.

Now revocation is the easy part; the harder part is 'reprovisioning' the cards into the new phone I buy the day after the mugging. If a bad man can cancel my wallet, that's a nuisance; if he can social-engineer a call centre into transferring my credit cards to a phone he controls, that's serious.

Here the phone company definitely does have an incentive to cooperate. In the typical case I will go into a shop operated by one of its partners and select a new phone. I then have to show ID, perhaps pass a credit check if I'm extending a contract, and wait while the sales clerk goes online to link my new phone and new SIM card to my old phone number. There is an interesting control point here: if I'm switching from one phone company to another, I typically have to go to the old company and get a code to release my phone number. This adds a few days' delay to the process; the switching cost thus introduced increases the phone company's lockin enough to be of value (while not so much as to seriously impair competition). Economic theory predicts that in service industries the value of a firm is equal to the total lockin of all the customers [7].

So my second suggestion is that we make social-engineering attacks harder and simultaneously reward phone companies that agree to participate fully in the system as follows. If the customer wishes to switch away from a participating phone company, then the scheme operator (whether the wallet provider or the TSM) should send her an unlock code by physical mail at her registered address. However a participating company should be able to reprovision its own customer on the spot (subject to joint approval by the relevant bank). Card portability, like number portability, should add just enough lockin for the phone companies to participate, but not so much as to strangle competition.

It would probably be unreasonable to let a phone company reprovision credit cards to a new phone on its own. The sales clerk's commission provides a perverse incentive; the phone company won't want all the liability; and in any case, when the customer buys a new phone for the first time, she has to interact with her bank (or at least with a TSM acting on its behalf) in order to load her first credit card. So here the proposal is that the bank that wants its card to be the default on the new phone must interact with the customer (or pay a TSM to do so).

The final missing piece in the puzzle is what incentive the other card providers might have to allow their own cards to be reissued by the lead bank. I suggest that there might be an industry scheme with uniform and non-discriminatory rules, perhaps to the effect that when the phone company and the lead bank authorise the reprovisioning of a phone, then all cards issued by members of the scheme should be re-enabled together. It could be both complex and invidious if the lead bank could selectively disable the cards of other banks that it considered to be acute competitors. There should also be an agreed level of indemnity and compensation in the event of reprovisioning errors.

Commercial terms are clearly for industry negotiation; but the wallet suppliers have a strong interest in encouraging agreement, and governments seeking to solve the problem of 'e-ID' may also have a role to play in bringing the various stakeholders to the table.

Conclusion

Federated authentication has mostly failed to work because the incentives were wrong. Identity providers assumed no liability and were open to traceless coercion; relying parties gained little benefit and had to cope with increased complexity; users rightly feared single points of failure.

Mobile wallets are both a problem and an opportunity. In the one market where NFC payments have already been deployed – Japan – things are somewhat fragmented. Not all customers can remotely disable lost or stolen phones; those who can't, have to call all their card providers, who have widely varying provisions for blocking and recovery. As a result, NFC payments there are largely limited to low-value prepaid cards. It is in the interests of all stakeholders to get a better outcome as NFC is deployed globally.

In this paper I argue that while wallets make the revocation part of federated authentication much more important, they may also make it more solvable – because the incentives of banks and phone companies are reasonably well aligned with the underlying customer contact problem. Banks do indeed wish to revoke their customers' compromised payment credentials, while phone companies for their part do indeed wish to sell new phones to customers whose phones have been lost or stolen.

With only a modest amount of tinkering and tweaking, the wallet providers who set the rules for this ecology may be able to ensure that the existing call-centre and know-your-customer resources of the banks and phone companies are deployed constructively to solve the problem. The key, I suggest, is that the bank or other card issuer which values the customer relationship the most should have its card set to be the default, as a reward for being the first port of call; that switching away from a participating phone company should become slightly harder; and that banks which fail to participate will lose market share when their cards are no longer reprovisioned automatically.

Finally, a robust mobile payment ecosystem with serious incentives to keep phone credentials bound to the people authorised to use them can provide a firmer platform for all sorts of other authentication services. I predict that in five years' time we will no longer think of recovering from a stolen phone by using the email account attached to it; we'll think instead of the wallet as the source of ground truth to recover credentials for other systems. It's the one thing everyone's got a real incentive to defend. Perhaps even the ambitions of the German government will fade away – and a stolen *Personalausweis* will be recovered using the citizen's mobile wallet, rather than vice versa.

Indeed, rather than being monopoly providers of "identity", the proper role of government should be to set the rules within which a market for federated authentication services can flourish.

Acknowledgement

Many of the ideas in this paper were developed while I was on sabbatical and working for Google in January–February 2011. I'm grateful for discussions there with Rob von Behren, Jason Waddle, Justin Brickell, Mark Andrews, Kosuke Suzuki, Vinay Rao, Ken Thompson, Mike Burrows, Rich Cannings, Ben Laurie and Alma Whitten. Others who have given useful feedback include Johann Bezuidenhoudt, Don Norman, Joe Bonneau, Shailendra Fuloria and Hyoungshick Kim. Needless to say the opinions expressed herein are mine rather than Google's.

Bibliography

[1] 'National Strategy for Trusted Identities in Cyberspace', at http://www.nist.gov/nstic

[2] Chris Soghoian, "Caught in the Cloud: Privacy, Encryption, and Government Back Doors in the Web 2.0 Era", 8 J. on Telecomm. and High Tech. L. 359, at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1421553

[3] 'SPKI/SDSI Certificates', at http://world.std.com/~cme/html/spki.html

[4] Ross Anderson, Steven Murdoch, "Verified by Visa and MasterCard SecureCode – or, How Not to Design Authentication", at Financial Cryptography 2010, at http://www.cl.cam.ac.uk/~rja14/Papers/fc10vbvsecurecode.pdf

[5] San-Tsai Sun, Yazan Boshmaf, Kirstie Hawkey, Konstantin Beznosov, "A Billion Keys, but Few Locks: The Crisis of Web Single Sign-On", LERSSE-RefConfPaper-2010-006, at http://lersse-dl.ece.ubc.ca/record/244

[6] Alex Hudson, "Are Call Centres the Factories of the 21st Century?" BBC News, Mar 10 2011, at http://www.bbc.co.uk/news/magazine-12691704

[7] Carl Shapiro, Hal Varian, '*Information Rules*', Harvard Business School Press, 1998